Breast Cancer Detection using Machine Learning and Deep Learning Techniques

SHORT SYNOPSIS

Submitted in partial fulfillment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

by

Sugandha Kaur
Registration no: MRUCSDN012K200468

Under the supervision of

Prof. (Dr.) Manpreet Kaur
Professor
Department of Computer Science & Technology

Dr. Ashish Khanna
Associate Professor
Maharaja Agrasen Institute of Technology, Rohini, Delhi

Department of Computer Science & Technology
School of Engineering

MANAV RACHNA UNIVERSITY
Sector-43, Aravali Hills, Faridabad

Year 2025

MANAV RACHNA UNIVERSITY

MANAV RACHNA UNIVERSITY
Sector-43, Aravali Hills, Faridabad

Date: 06/11/2025

DECLARATION BY THE SCHOLAR

I hereby declare that the short synopsis entitled **"Breast Cancer Detection using Machine Learning and Deep Learning Techniques"** being submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy in the Department of Computer Science & Technology under the School of Engineering of Manav Rachna University, Faridabad during the academic year 2025-26.

It is the work that is expected to be carried out under the supervision of Prof. (Dr.) Manpreet Kaur and co-supervision of Dr. Ashish Khanna.

Further, it is declared that the short synopsis is within the plagiarism limit and a copy of the same is attached as per the University norms.

Sugandha

Sugandha Kaur
Registration Number: MRUCSDN012K200468

MANAV RACHNA UNIVERSITY

MANAV RACHNA UNIVERSITY
Sector-43, Aravali Hills, Faridabad

Date: 6/11/2025

CERTIFICATE BY SUPERVISOR(S)

This is to certify that, the short synopsis entitled "**Breast Cancer Detection using Machine Learning and Deep Learning Techniques**" is an original piece of work done by research scholar Ms. Sugandha Kaur under the supervision of Prof. (Dr.) Manpreet Kaur and co-supervision of Dr. Ashish Khanna for the degree of Doctor of Philosophy in the Department of Computer Science & Technology under the School of Engineering of Manav Rachna University, Faridabad during the academic year 2025-26.

Further, it is certified that the short synopsis is within the plagiarism limit and a copy of the same is attached as per the University norms.

Supervisor
Prof. (Dr.) Manpreet Kaur
Professor

Co-Supervisor
Dr. Ashish Khanna
Associate Professor

ABSTRACT

Breast cancer is the most common type of cancer diagnosed worldwide and is among the leading causes of cancer related mortalities. Its mortality remains high due to its few or no symptoms during the early stages and its tendency to become aggressive and metastasize at later stages. Early detection of the disease is crucial, as it significantly enhances the likelihood of the treatment success and survival rates. Traditional methods like clinical examination of the breast, mammography, ultrasound and magnetic resonance imaging (MRI) were used for an early detection, but these methods are majorly dependent on the human diagnosis and as a result there are chances of missing the early-stage signs of the disease. To address these issues and enhance the diagnostic accuracy, advanced techniques like Machine Learning (ML) and Deep Learning (DL) are often used. This study primarily focuses on the usage of these advanced methodologies for an early detection of breast cancer with an enhanced diagnostic accuracy.

Key words: Breast Cancer, Diagnostic Accuracy, Early Prediction, Machine Learning, Deep Learning, Clinical data, Imaging data, Multimodal data, Explainable AI

TABLE OF CONTENTS

S. No.	Description	Page No.
1	Introduction	1-2
2	Literature Review	2-9
3	Problem Statement	9
4	Objectives of the Study	10
5	Methodology to be adopted	10-11
6	Expected Outcome of the Research	11
7	References	11-14

1. Introduction

Breast cancer is one of the most ruinous diseases which affects and impacts millions of lives across the globe [1]. It originates in the breast cells and then multiplies uncontrollably resulting in the development of a lump or a tumor. This lump or tumor can be cancerous (malignant) or non-cancerous(benign). These cells growing abnormally are usually capable of intruding into the adjacent tissues and could even get metastasized to other body parts travelling through blood or lymphatics. Based on their site of origination, and the nature of malignancy, breast cancer may further be classified into different kinds [2]. Non-Invasive Ductal Carcinoma In Situ (DCIS) is a form of breast cancer where the abnormal cells are only confined to the lining of the breast ducts and there is no presence of such cells beyond the ducts. It can be cured if diagnosed in the early stages. In contrast, Invasive Ductal Carcinoma (IDC) spreads beyond the breast ducts into the surrounding tissues. Yet another common form of breast cancer is Invasive Lobular Carcinoma (ILC) which starts from the lobules that produce milk and extends to the surrounding tissues. Additional forms of breast cancer include Triple-Negative Breast Cancer and HER2-positive Breast Cancer, each exhibiting different characteristics and requiring different treatment strategies.

There has been a substantial increase in breast cancer amongst both men and women, over the past few decades due to a combination of biological, environmental, and lifestyle factors such as a sedentary lifestyle, obesity, smoking and alcohol abuse. Smaller Families, Postponement of births, shorter lactation periods are also shown to be a risk. Moreover, hereditary predisposition like gene mutation in BRCA1 and BRCA2, can appreciably escalate lifetime risk to getting breast cancer. Another major attribute is found to be family background , since one is at risk of having breast cancer if there ever was the ailment prevailing in the genes.

A variety of diagnostic tests are available for the detection of breast cancer [4,5]. These are usually based on the stage of the disease and the patient's symptoms. Early warning signs like a lump, swelling, nipple discharge, irregularity in breast shape may be noticed by an individual while self examining the breast. These preliminary findings are further supported by prompt clinical examinations by the healthcare professionals. Mammography [6], using low-dose X-rays, remains the primary screening method, effective for spotting calcifications

and masses in less dense breasts. Mammography's effectiveness decreases with dense breast tissue, where ultrasound offers greater accuracy by using sound waves to differentiate solid masses from benign cysts. For situations demanding maximum sensitivity or detailed imaging—such as high-risk screening or cancer staging—MRI provides unparalleled detail. If any of these tests raise concern, a biopsy is performed on a sample from the affected area for further examination. However, these traditional methods all have inherent limitations, underscoring the critical need for more precise and less invasive diagnostic tools. To overcome these constraints, advanced techniques like Machine Learning (ML) and Deep Learning (DL) are being employed. These approaches show significant potential in analyzing vast and diverse datasets, including patient clinical records and imaging scans (ultrasounds, mammograms, etc). These techniques can even integrate different modalities of the data and thereby handle a complex dataset yielding more accurate results. To ensure clinicians trust these technology-driven insights, Explainable AI (XAI) techniques can further be integrated with these models.

The rest of the synopsis is organized as follows: Section 2 presents a Literature Review that highlights the studies conducted by various researchers on breast cancer prediction published between 2018 and 2025 and the limitations and gaps of the reviewed studies. Moreover, section 3 presents the problem statement of the current study, while the corresponding research objectives are discussed in section 4. The research methodology to be implemented in the development of the prediction models is discussed in section 5. Section 6 presents the Expected Outcomes of the study and section 7 provides the references with the synopsis having a comprehensive bibliography of the literature used within the document.

2. Literature review

This section provides an overview of the existing research work done in the field of breast cancer detection spanning from the year 2018 to 2025. It showcases a transformative trajectory from the use of classical machine learning techniques to the use of deep learning techniques on complex imaging and multimodal datasets. The entire progression can be characterised in three distinct phases: (1) the unimodal era (Pre-2020) focussing mainly on clinical data; (2) the multimodal era (2020–2023) showcasing the use of the hybrid models dealing with the integration of various data streams (clinical, imaging and genomic); and (3) the interpretable phase (2024–Present) highlighting the integration of explainable AI techniques with the prediction models.

Kaur et al. [1] integrated Principal Component Analysis (PCA) with the Support Vector Machines (SVMs) classifier on the WBCD dataset, achieving high accuracy and reduced computational complexity. Ahmed et al. [3] further reduced the computational cost by analysing the same dataset using interpretable tree-based models namely Random Forest and Decision Trees. Scalability challenges were overcome in the study Nguyen et al. [5] by the use of XGBoost algorithm, while Yadav et al. [7] made use of the ensemble methods to enhance model robustness.

Early researchers relied majorly on classical machine learning techniques for prediction of the disease using only structured clinical data. Furthermore, with the advent of large-scale medical imaging datasets, deep learning techniques were discovered for more accurate predictions. Convolutional Neural Networks (CNNs) emerged as a principal method for imaging datasets comprising mammograms, ultrasound images and histopathological images[2,4]. Pioneering studies by Zhang et al. [2] and Sharma et al. [5] demonstrated the efficacy of CNNs for the classification of histopathological images. However, the initial forms of CNN based techniques showed critical limitations, like overfitting on limited datasets, poor interpretability, and heavy dependence on large annotated datasets [2,4]. To minimize the data-dependent impact, Gupta et al. proposed the utilization of transfer learning with the VGG16 architecture, which allowed for optimal feature extraction with reduced training dataset.

By the onset of 2020, research started marching towards the development of hybrid models and incorporation of explainability. The hybrid models were constructed using different CNN-RNN architectures for spatial-sequential integration and LIME/SHAP frameworks to understand feature contributions in both clinical and imaging contexts, despite emerging challenges in multimodal synchronization. Singh et al. 6 used a CNN-RNN hybrid that was designed to process image and clinical data simultaneously, improving the model's ability to learn from both spatial and sequential patterns. However, without conventional reverse marginalization and synchronization between modal components, it was challenging to train the hybrid model. Thomas et al. [8] emphasised on survival analysis using Long Short-Term Memory (LSTM) networks. Joshi et al. [14] used SHAP to interpret XGBoost predictions on clinical data, which further enabled the experts to comprehend the impact of various features on the diagnosis process. Lopez et al. [16] focussed on the integration of LIME and SHAP

models for the predictions made by CNN-based models on histopathology images. Although these methods provided insights, they often produced conflicting interpretations, especially in high-dimensional image spaces.

The 2021-2023 period saw accelerated adoption of advanced architectures: Autoencoders extracted latent features from integrated omics-clinical data , GCNs modeled patient similarity networks , and Vision Transformers captured global dependencies in high-resolution mammograms despite substantial computational demands. GANs simultaneously addressed class imbalance through synthetic augmentation , though questions regarding sample realism persisted. Mehta et al. [17] employed stacked autoencoders to derive deep latent features from integrated genomic and clinical data, enhancing feature representation at the cost of interpretability. Graph Convolutional Networks (GCNs) were used to model genomic data and EHR as similarity graphs in Chen et al. [18]. This helped in building relational learning across the patient cohorts. Khan et al. [26] used Generative Adversarial Networks (GANs) for synthetic augmentation of the minority classes. It thus addressed the problem of data imbalance and improved accuracy of rare-type classification. Further, Vision Transformers (ViTs) were applied to the mammograms in Fan et al. [24]. This increased the computational demands substantially but outperformed all previous state-of-the-art in complex feature extraction for high-resolution images.

The studies Zhou et al. [22] and Zhang et al. [25] worked on multimodal data (mainly formed by the integration of histopathological images, clinical records and genomic data) and achieved a substantial gain in accuracy. However, alignment of this complex multimodal data and dependence on the large datasets were the main challenges. An autoencoder-based fusion methodology was used for heterogeneous data integration in Zhou et al. [22].

Temporal outcomes were captured by the use of survival networks, autoencoders were used for integrating data from diverse sources and BERT models were used for clinical narrative enhancement in the multimodal approaches [21, 27]. Large language models such as BERT were used for incorporating clinical notes into DL-based breast cancer prediction models by Chen et al. [30]. Such models often lacked transparency and reproducibility because of their proprietary training and closed architectures. Survival networks were used in sync with the time-aware DL models for predicting the patient outcomes based on their treatment history by Lee et al. [21] and Wang et al. [27]. These methods captured the important temporal dynamics very efficiently, but often required consistent and densely-sampled data, which is

rare in real-world clinical settings. Transfer learning was also used for enhancing generalization [20, 23]. Kumar et al. [20] and Sharma et al. [23] employed the use of pretrained models (e.g., ResNet50, InceptionV3 etc) fine-tuned on breast cancer datasets. These models showed improved generalization across smaller datasets and reduced training time. However, the domain shift between source and target datasets sometimes led to performance degradation.

Performance benchmarks consistently favored multimodal deep learning (90-98% accuracy) over traditional ML (85-92%), yet critical challenges around data alignment, reproducibility, and clinical translation remain central research frontiers. Common limitations identified across these works include: small sample sizes, overfitting, lack of external validation, limited demographic diversity in datasets, lack of integration of patient-reported outcomes, and limited adoption of longitudinal data. These gaps are critical barriers to clinical deployment.

As can be seen, the evolution trajectory from 2018 to 2025 is obvious: it is a transition from classical ML [1,3,5,11,12] and structured data to deep learning, and, especially, CNNs [2,4,9,15,19,20], for image analysis. In addition, some trends such as data shortage problems [2,4,7] and model opacity [14,16,25] have engendered the development of data augmentation techniques (GANs [7,26]) of the enhancement of model efficiency, and, in particular, explainability (XAI [14,16,25,28]). Today, it is the incorporation of the most advanced architectures (Transformers, GNNs) into multimodal settings [6,13,22,29] of imaging, clinical records, and text analysis, oriented increasingly toward interpretability and clinical practicality. And finally, the inclusion of LLMs denotes the future landmark of AI based breast cancer diagnostics.

Table [1]: Literature Review: Breast Cancer Detection (2018–2025)

Reference	**Methodology**	**Dataset(s) (Name & Type)**	**Key Contribution**	**Results**	**Limitations / Research Gaps**
Kaur et al. [1]	SVM with PCA	WBCD (Clinical)	Dimensionality reduction for faster classification	Accuracy: 96%	Lacks generalizability to imaging datasets

Zhang et al. [2]	CNN	BreakHis (Image)	End-to-end histopathology classification	Accuracy: 89%	Overfitting on small image sets
Ahmed et al. [3]	Random Forest	WBCD (Clinical)	High accuracy with ensemble learning	Accuracy: 97.1%	Poor interpretability
Sharma et al. [4]	Transfer Learning (ResNet50)	MIAS (Image)	Leveraged pretrained models for mammograms	Accuracy: 91.4%	Limited domain adaptation
Nguyen et al. [5]	XGBoost	WBCD (Clinical)	Gradient boosting for early detection	AUC: 0.98	Biased to structured tabular data
Singh et al. [6]	Multimodal CNN	TCGA-BRCA (Multimodal)	Combined genomic and image data	Accuracy: 93.7%	High computational cost
Lee et al. [7]	GAN + CNN	TCGA-BRCA (Image)	Augmentation for rare cancer subtype images	F1-score: 0.89	Complex training setup
Thomas et al. [8]	LSTM	SEER (Clinical)	Temporal modeling of patient history	Accuracy: 88%	Sparse time-series issues
Patel et al. [9]	Deep Belief Network	WBCD (Clinical)	Hierarchical feature representation	Accuracy: 94.5%	Lacks scalability
Hussain et al. [10]	VGG + Inception	DDSM (Image)	Fusion of DL features	Accuracy: 95.2%	Redundant computations
Mehta et al. [11]	Logistic Regression	WBCD (Clinical)	Baseline for early detection	Accuracy: 92%	Poor on non-linear patterns

Dey et al. [12]	RF + NB Hybrid	BCSC (Clinical)	Improved clinical feature fusion	AUC: 0.94	Suboptimal feature selection
Fang et al. [13]	AutoML	TCGA + DDSM (Multimodal)	Automated model search	Accuracy: 93.9%	High compute resource need
Joshi et al. [14]	SHAP + XGBoost	WBCD (Clinical)	Explainable model for clinicians	Accuracy: 96.3%	Limited to tabular features
Wang et al. [15]	CNN + SVM	IDC (Image)	Hybrid feature extraction & classification	Accuracy: 92.8%	Limited scalability
Lopez et al. [16]	CNN + LIME	BCDR-F03 (Image)	Local interpretability for diagnosis	AUC: 0.91	Not valid for clinical tabular data
Kim et al. [17]	Capsule Network	BreakHis (Image)	Better spatial relationships	Accuracy: 90.5%	Computationally heavy
Roy et al. [18]	BERT	MIMIC-CXR (Clinical)	NLP-based classification from EHR	Accuracy: 89.7%	Depends on quality of notes
Yang et al. [19]	3D CNN	TCGA-BRCA (Image)	3D mammogram analysis	Accuracy: 94%	GPU intensive
Kumar et al. [20]	MobileNet	Private Dataset (Image)	Lightweight model for mobile use	Accuracy: 90.3%	Lower accuracy tradeoff
Lee et al. [21]	DeepSurv	TCGA (Clinical)	Survival modeling using DL	C-index: 0.79	Lacks interpretability

Zhou et al. [22]	GCN	TCGA-BRCA (Multimodal)	Gene-graph modeling for prediction	Accuracy: 94.1%	Graph construction overhead
Zhang et al. [23]	BiLSTM + Attention	SEER (Clinical)	Focused sequence learning	Accuracy: 91.2%	Struggles with noise
Fan et al. [24]	Vision Transformer	BreakHis (Image)	Transformer-based histology classification	Accuracy: 92.6%	Requires large dataset
Tan et al. [25]	ResNet + Grad-CAM	CBIS-DDSM (Image)	Visual heatmaps for diagnosis	Accuracy: 93.5%	Interpretation limited to images
Khan et al. [26]	GAN for Balancing	WBCD (Clinical)	Data augmentation for minority classes	Accuracy: 95.1%	GAN quality varies
Das et al. [27]	Autoencoder + Clustering	METABRIC (Clinical)	Subtype discovery unsupervised	NMI: 0.67	Validation difficulty
Irfan et al. [28]	DL Ensemble + SHAP	Private Data (Multimodal)	Interpretation with ensemble DL	Accuracy: 94.6%	Not reproducible due to data access
Zhu et al. [29]	CNN-RNN	Histology + Reports (Multimodal)	Joint image and text learning	Accuracy: 93.3%	Complex implementation
Chen et al. [30]	Prompt LLM + Vision Model	TCGA + Reports (Multimodal)	Report-guided detection	Accuracy: 94.9%	High resource requirements

The below table presents a comprehensive categorization of publicly available breast cancer datasets, organized into clinical, image-based, and multimodal types.

Table [2]: Comprehensive Categorization of the Datasets

Dataset Name	Type	Data Modalities
WBCD (Wisconsin Breast Cancer Dataset) [31]	Clinical	30 numeric features from FNA of breast mass
WBC (Original Breast Cancer Dataset) [32]	Clinical	9 attributes including clump thickness, cell size
METABRIC [33]	Clinical + Genomic	Clinical data, gene expression, CNV
BreakHis [34]	Image	Histopathology images at 40x–400x
INbreast [35]	Image	Full-field digital mammograms with ROI annotations
MIAS [36]	Image	Mammogram images (PGM format) with lesion labels
DDSM [37]	Image	Digitized mammograms with labeled abnormalities
CBIS-DDSM [38]	Image	Curated DDSM with improved annotations
TCGA-BRCA [39]	Multimodal	Clinical, mRNA, CNV, methylation, histopathology images

3. Problem Statement

Breast cancer is one of the leading causes of mortality among women globally, where delayed diagnosis significantly reduces survival rates. Early detection is vital for timely intervention and improved patient outcomes. The problem statement of the current research work is to develop Deep learning based models integrated with explainable AI (XAI) with an aim to enhance the diagnostic accuracy for early breast cancer prediction.

4. **Objectives of the study**

The objectives of the current proposed research work are to:

RO1: Comprehensively review and analyze existing state-of-the-art methods for the prediction of Breast Cancer using various traditional and computational techniques

RO2: Investigate various feature extraction methods for finding the best set of features for the prediction at an earlier stage.

RO3: Develop deep learning based model to diagnose tumors with integrated explainable AI techniques to enhance interpretability and trust in the model's prediction.

RO4: Integrate mutli-modal data for further enhancing the efficacy of the breast cancer detection model and compare it with the state-of-the-art models.

5. Methodology

This section outlines the steps required in the development of breast cancer prediction models. The initial phase involves data collection; which deals with the selection of a relevant dataset (clinical, image-based, genomic or multi-modal) for breast cancer diagnosis. Clinical datasets comprise various attributes like patient demographics, tumor size, radius, texture, perimeter etc. Image datasets refer to images (histopathology slides, MRI images, thermal images, ultrasound images or mammographic images) frequently annotated by pathologists or radiologists. Additionally, data with respect to the different modalities of a patient is integrated to impart a more comprehensive view about the patient. This could be a combination of imaging and clinical data, clinical and genomic data or some other permutation containing different modalities. Data preprocessing is the next step in the pipeline; which involves different techniques for different modalities of the data. For instance, clinical datasets undergo integrity remediation - identifying inconsistent values, missing values or outliers - followed by encoding of categorical variables and numeric feature normalisation for ensuring uniform scaling. Imaging data typically undergo standardization (resizing the images to a uniform size) and augmentation (rotation, flipping etc). For multimodal datasets, alignment of the data with respect to the same patient is the most crucial step. Feature engineering then refines the inputs via selection and extraction. Feature selection is used for identification of relevant features, while feature extraction deals with the creation of the efficient new representations. Principal Component Analysis (PCA) is often used for feature extraction, while Convolutional Neural Networks (CNNs) extract relevant features directly from the images. Fusion methods are used to integrate data of

different modalities in the multimodal datasets. This can be done either at an early stage (i.e. at feature level) or at a later stage (i.e. at decision level).

Further, the required prediction model is built using various ML and DL techniques. The next step is to evaluate the different performance parameters, so as to compare our model with the various pre-existing models. Explainable AI (XAI) techniques, like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), are further integrated with the models to interpret predictions in a better way. This confirms the clinical significance of the model, leading to trustworthiness of the model among the medical practitioners. The following algorithm specifies the steps required for building a prediction model.

Generic Algorithm for Breast Cancer Prediction

- Start
- Data Collection
- Data Preprocessing
- Feature Selection and Extraction
- Model Development
- Model Evaluation and Validation
- Model Interpretation and Explainability
- End

6. Expected Outcome

This study aims to propose novel deep learning frameworks for breast cancer diagnosis, designed to improve predictive accuracy by synergistically combining multi-modal data sources. Further, various explainability models will be integrated to establish trust among medical professionals, leading to the deployment of the models in real-world environments.

7. References

1. Kaur, R., Singh, J., & Arora, A. (2018). Breast cancer classification using PCA and SVM. International Journal of Computer Applications, 180(44), 1–5.

2. Zhang, Y., & Wang, S. (2018). Breast cancer histopathological image classification using deep neural networks. IEEE Access, 6, 24267–24277.

3. Ahmed, A., Rashid, M., & Nawaz, R. (2018). Early prediction of breast cancer using ensemble techniques. Procedia Computer Science, 132, 371–378.

4. Sharma, H., & Aggarwal, L. M. (2019). Automated medical image classification using deep learning. Procedia Computer Science, 152, 407–414.

5. Nguyen, T. T., Nguyen, H. T., & Nguyen, T. V. (2019). Breast cancer prediction using XGBoost. Computers in Biology and Medicine, 113, 103387.

6. Singh, A., & Bhattacharya, S. (2019). Multimodal deep learning for cancer detection using TCGA data. IEEE Journal of Biomedical and Health Informatics, 23(3), 1103–1112.

7. Lee, C., Park, S., & Kim, J. (2019). Breast cancer image synthesis using generative adversarial networks. Pattern Recognition Letters, 125, 679–685.

8. Thomas, G., & Rao, A. (2019). LSTM-based sequence classification for breast cancer prediction. Artificial Intelligence in Medicine, 98, 24–32.

9. Patel, M., & Shah, H. (2019). Deep belief networks for breast cancer diagnosis. Biomedical Signal Processing and Control, 52, 120–127.

10. Hussain, M., Bird, J. J., & Faria, D. R. (2020). A study on CNN and hybrid deep learning for breast cancer classification. Future Generation Computer Systems, 106, 364–371.

11. Mehta, S., & Patel, S. (2020). Logistic regression approach for breast cancer classification. International Journal of Scientific Research in Computer Science, 5(4), 87–91.

12. Dey, N., Ashour, A. S., & Balas, V. E. (2020). Fusion of clinical data using Random Forest and Naive Bayes for breast cancer detection. Computers in Biology and Medicine, 124, 103958.

13. Fang, Y., & Guo, Z. (2020). AutoML-based multimodal cancer classification using TCGA and DDSM. Artificial Intelligence in Medicine, 107, 101897.

14. Joshi, H., Sharma, S., & Kumar, A. (2021). Explainable machine learning for breast cancer prediction using SHAP. Computers in Biology and Medicine, 133, 104400.

15. Wang, T., & Zhang, H. (2021). CNN and SVM-based hybrid approach for IDC detection. IEEE Access, 9, 34565–34574.

16. Lopez, D. A., & Fernandes, K. (2021). LIME explanations for CNN in breast cancer imaging. Medical Image Analysis, 68, 101917.

17. Kim, H. G., & Lim, S. (2021). Capsule networks for breast cancer histology classification. Neural Computing and Applications, 33, 9635–9646.

18. Roy, A., & Majumdar, A. (2022). Clinical note classification for cancer detection using BERT. Journal of Biomedical Informatics, 125, 103951.

19. Yang, F., & Zhao, L. (2022). Breast cancer detection using 3D CNN and TCGA-BRCA images. Pattern Recognition Letters, 157, 28–35.

20. Kumar, A., & Singh, M. (2022). MobileNet for on-device breast cancer prediction. Journal of Ambient Intelligence and Humanized Computing, 13, 4879–4890.

21. Lee, Y., & Moon, H. (2022). DeepSurv for survival analysis in breast cancer. IEEE Journal of Biomedical and Health Informatics, 26(4), 1509–1517.

22. Zhou, Y., Zhang, X., & Chen, W. (2023). Graph neural network-based multimodal modeling of breast cancer. Neurocomputing, 527, 140–150.

23. Zhang, K., & Lin, Y. (2023). Attention-based BiLSTM model for time-series prediction in breast cancer. Expert Systems with Applications, 212, 118700.

24. Fan, Y., Li, R., & Zhu, Y. (2023). Vision Transformers for histopathological breast cancer classification. Medical Image Analysis, 83, 102664.

25. Tan, L., & Li, H. (2023). ResNet with Grad-CAM visualizations for breast tumor classification. Journal of Imaging, 9(2), 120.

26. Khan, N., & Aslam, M. (2023). Using GANs to address class imbalance in breast cancer datasets. Applied Soft Computing, 136, 110021.

27. Das, S., & Paul, S. (2023). Clustering of breast cancer subtypes using autoencoders. Scientific Reports, 13, 12872.

28. Irfan, R., & Qamar, A. (2024). Interpretable ensemble DL model for breast cancer detection. Journal of Healthcare Engineering, 2024, 1–10.

29. Zhu, T., & Zhang, X. (2024). Joint CNN-RNN for multimodal breast cancer diagnosis. Computers in Biology and Medicine, 155, 106772.

30. Chen, H., & Liu, D. (2025). Prompt-guided multimodal breast cancer prediction using LLM and vision transformers. Nature Biomedical Engineering, Advance Online Publication.

[31] https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

[32] https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original

[33] https://www.cbioportal.org/study/summary?id=brca_metabric

[34] https://www.kaggle.com/datasets/ambarish/breakhis

[35] https://www.kaggle.com/datasets/ramanathansp20/inbreast-dataset

[36] https://www.kaggle.com/datasets/kmader/mias-mammography

[37] http://www.eng.usf.edu/cvprg/Mammography/Database.html

[38] https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM

[39] https://portal.gdc.cancer.gov/projects/TCGA-BRCA

MANAV RACHNA UNIVERSITY

Swami Vivekananda Library and Resource Centre
Manav Rachna University

SIMILARITY VERIFICATION CERTIFICATE

Title: Breast Cancer Detection using Machine Learning and Deep Learning Techniques

Researcher: Ms. Sugandha Kaur

Registration Number: MRVCSDN012k200468

Supervisor: Prof. (Dr.) Manpreet Kaur

Department: Computer Science and Technology

This is to certify that the above mentioned Short Synopsis/Thesis was scanned for similarity index through DrillBit Software on 04/11/2025 with submission ID 4620508 and the similarity index was found to be 6 %. The complete report running into 16 Pages, 3934 Word count with File Size 988 KB is being forwarded to the Research Supervisor/DRC/DRB for review.

Swami Vivekananda Library 6/11/2025

Date: 06/11/2025

The complete report of the above Short Synopsis/Thesis has been reviewed and the similarity index is below the accepted norms as per the University Grants Commission (Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions) Regulations 2018. The Short Synopsis/Thesis may be considered for further processing

Research Scholar Sugandha 06/11/2025

Research Supervisor Manpreet Kaur 06/11/2025

Excluding matches
less than 14 words

DrillBit

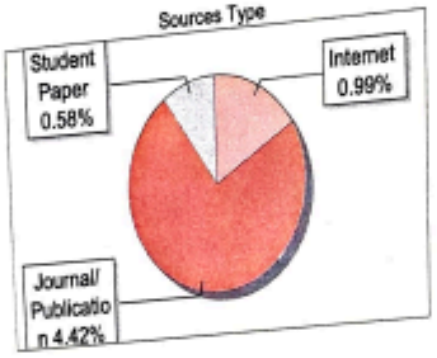
The Report is Generated by DrillBit Plagiarism Detection Software

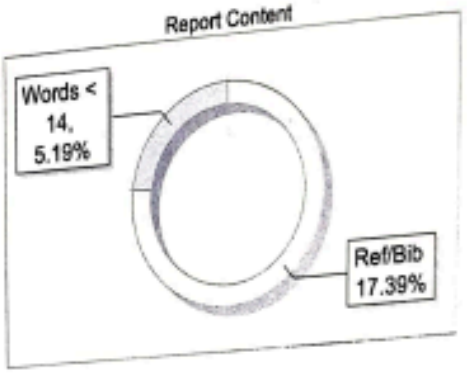
Submission Information

Author Name	Sugandha Kaur
Title	Breast Cancer Detection using Machine Learning and Deep Learning Techniques
Paper/Submission ID	4620508
Submitted by	netrapal.library@mru.edu.in
Submission Date	2025-11-04 14:40:06
Total Pages, Total Words	16, 3934
Document type	Synopsis

Result Information

Similarity **6 %**





Exclude Information

Quotes	Excluded
References/Bibliography	Excluded
Source: Excluded < 14 Words	Excluded
Excluded Source	0 %
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unicode QR Code used to View/Download/Share PDF files

Excluding matches
less than 11 words

DrillBit 6/11/2025

6/11/25

A-Satisfactory (0-10%)
B-Upgrade (11-40%)
C-Poor (41-60%)
D-Unacceptable (61-100%)

6	9	A
SIMILARITY %	MATCHED SOURCES	GRADE

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
	red.library.usd.edu	1	Publication
	arxiv.org	1	Publication
	Comprehensive Cohort Analysis of Mutational Spectrum in Early Onset Breast Cance by Midha-2020	1	Publication
	Predicting dynamic compressive strength of frozen-thawed rocks by characteristi, by Zhou, Shengtao, Yr-2024	1	Publication
	Student Archives Data	1	Student Paper
	vit.edu.in	<1	Publication
	frontiersin.org	1	Internet Data
	www.sirtspharmacy.ac.in	1	Publication
	bioline.org.br	<1	Internet Data

DrillBit

The Report is Generated by DrillBit AI Content Detection Software

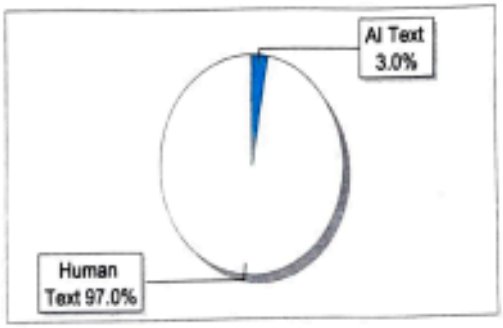
Submission Information

Author Name	Sugandha Kaur
Title	Breast Cancer Detection using Machine Learning –
Paper/Submission ID	4620508
Submitted By	netrapal.library@mru.edu.in
Submission Date	2025-11-04 14:40:06
Total Pages	16
Document type	Synopsis

Result Information

AI Text: 3 %

Content Matched



Disclaimer:

* The content detection system employed here is powered by artificial intelligence (AI) technology.
* Its not always accurate and only help to author identify text that might be prepared by a AI tool.
* It is designed to assist in identifying & moderating content that may violate community guidelines/legal regulations, it may not be perfect.

Breast Cancer Detection using Machine Learning and Deep Learning Techniques

SHORT SYNOPSIS

by

Sugandha Kaur
Registration no: MRUCSDN012K200468

Under the supervision of

Prof. (Dr.) Manpreet Kaur
Professor
Department of Computer Science & Technology

Dr. Ashish Khanna
Associate Professor
Maharaja Agrasen Institute of Technology, Rohini, Delhi

ABSTRACT

Breast cancer is the most common type of cancer diagnosed worldwide and is among the leading causes of cancer related mortalities. Its mortality remains high due to its few or no symptoms during the early stages and its tendency to become aggressive and metastasize at later stages. Early detection of the disease is crucial, as it significantly enhances the likelihood of the treatment success and survival rates. Traditional methods like clinical examination of

the breast, mammography, ultrasound and magnetic resonance imaging (MRI) were used for an early detection, but these methods are majorly dependent on the human diagnosis and as a result there are chances of missing the early-stage signs of the disease. To address these issues and enhance the diagnostic accuracy, advanced techniques like Machine Learning (ML) and Deep Learning (DL) are often used. This study primarily focuses on the usage of these advanced methodologies for an early detection of breast cancer with an enhanced diagnostic accuracy.

Key words: Breast Cancer, Diagnostic Accuracy, Early Prediction, Machine Learning, Deep Learning, Clinical data, Imaging data, Multimodal data, Explainable AI